DOCUMENT RESUME

ED 210 275 TM 810 716

AUTHOR Halpin, Glennelle; Halpin, Gerald
TITLE Testing: A Key to High Student Achievement but Low Student Ratings?
PUB DATE Sep 80
NOTE 26p.; Revision of a Paper presented at the Annual Meeting of the American Psychological Association (Montreal, Canada, September, 1980).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Academic Achievement; Course Evaluation; Difficulty Level; Higher Education; *Learning Processes; Literature Reviews; *Retention (Psychology); *Student Attitudes; *Study; Test Format; *Testing; *Test Use; Undergraduate Students

ABSTRACT

A review of the literature on the influence of testing on learning and retention reveals the need for more comprehensive research findings in this regard. This study was designed to investigate further the direct effects of tests in contrast with no tests on learning and retention in ongoing college-level classes with both the instructor and the students going about the daily business of teaching and learning. It probed further to determine if it was taking the test or studying for it or both which effected learning and retention if indeed such effects were replicable. Test type (multiple-choice and short answer) and item complexity (knowledge and concept) were also variables studied. Moving from the cognitive domain to the affective domain, this study focused on the students' feelings when they did/did not have to study for and take a test. Analyses of the resulting data showed that students who studied for and took a test not only achieved more but also retained their learning longer than students who "studied in order to learn than for a test." However, student ratings of construction were lower when students were tested. (Author/AL)

Testing: A Key to High Student Achievement but Low Student Ratings?

Glennelle Halpin and Gerald Halpin
Auburn University

Revision of a paper presented at the meeting of the American Psychological Association, Montreal, Canada, September 1980.

Tests: A Key to Student Learning and Retention But Low Student Ratings of Instruction?

For several decades there has been much theoretical discussion of the influence of testing on learning and retention. A number of research studies have been conducted to determine the extent and the nature of this influence. Over 45 years ago, White (1932) reviewed the literature on testing as an aid to learning and noted that the exam has "always" been regarded as a valuable instrument for motivating learning. However, in a review years later of the influence of the evaluating instrument on students' learning and retention, Balch (1964) commented that there had been much theoretical discussion of the value of testing in the learning process but only spasmodic research in the area. "Unfortunately," he concluded, "the number of experimental studies seems to be decreasing, few definitive answers are available, and there is apparently no comprehensive research findings upon which to base further studies" (p. 169). One definitive answer that seems to evolve from the research reviewed by Balch is that the use of an evaluation instrument (as opposed to not using one) influences student learning. However, he clearly pointed out that in the studies reviewed, "investigators sometimes allowed unconfirmed theory to color their conclusions" (p. 177), and he called for more research with better experimental design and a more careful analysis of results.

Studies since Balch's review tend to support the use of testing to promote student learning. Gaynor and Millham (1976), for example, found that academic performance of students in an introductory psychology course differed significantly as a function of the frequency of the examinations. Students who received weekly tests answered significantly more questions than did those students who received only midterm and final examinations. Likewise, Fisher,

Williams, and Roth (Note 1) found that undergraduate students tested weekly in an upper division science course outperformed those tested at midterm and end of term. On a retention test given 2 years later, the more frequently tested group scored 8 percentage points higher than the less frequently tested group, a strongly suggestive ($p < .10$) though not significant difference. A body of literature (cf. Landauer & Ainslie, 1975; LaPorte & Voss, 1975; Anderson, Surber, Biddle, Zych, & Lieberman, Note 2) supports this conjecture that tests are effective in reducing the rate of decay of knowledge and skills.

The trend in these more recent studies seems to be toward accepting testing as an important variable in learning and retention. Whether or not to test seems to have been replaced by such questions as those regarding the timing of quizzes as in the Gaynor and Millham (1976) study and in the Fisher et al. (Note 1) study. Test mode and test complexity have also been variables of interest (Voss, 1974; Fisher et al., Note 1; Anderson et al., Note 2; Sanjivamurthy & Kumar, Note 3). Even test-like events are capturing a share of the research as the effects of adjunct questions in written and oral presentations are studied (Frase, 1970; Koran & Koran, 1975; Rothkopf, 1972; Sanders, 1973; McKenzie & Schadler, Note 4).

The exclusion as a viable research question of whether or not to test may be a premature one, though, based on inadequate research or on conclusions based on unconfirmed theory as Balch (1964) noted. In a study of retention as a function of type and complexity of a test in an ongoing graduate-level educational psychology class, Halpin, Halpin, and Harrington (Note 5) found no significant differences in retention for groups given as treatment a multiple-choice test, a short-answer test, or no test. These results are contradictory to the generally accepted notion that a test is more influential on learning

than no test and seem to indicate that the issue is far from being conclusive. More research seems to be needed. This need for more research on the direct use of tests as treatment was also reflected by Anderson et al. (Note 2) when they said of their work:

> Two experiments were completed whose purpose was to investigate the direct effect of questioning. By a "direct effect" we mean the increment in performance which is observed when a question asked during or shortly after exposure to text is repeated on a later test. The direct effect of questioning is invariably larger than the indirect effect which has captured the lion's share of attention from the research community since the work of Rothkopf" (p. 3).

This study, then, was designed to investigate further the direct effects of tests in contrast with no tests on learning and retention in an ongoing class with both the instructor and the students going about the day by day business of teaching and learning. It probed further to try to determine if it was taking the test or studying for the test or both which effected learning and retention if indeed such effects were replicable. Test type (multiple-choice and short answer) and item complexity (knowledge and concept) were also variables studied since prior relevant research is inconclusive (Meyer, 1934, 1935; Vallance, 1947; Hakstian, 1971; Fisher et al., Note 1; Sanjivamurthy et al., Note 3). Finally moving from the cognitive domain to the affective domain, this study focused on the students' feeling when they did/did not have to study for and take a test.

More specifically, objectives for this study were:

To see if students in a regular classroom setting who expect and study for a test achieve on a classroom examination at the same level as students who do not expect a test but instead study "in order to learn."

To determine if studying for a test, actually taking a test, or a combination of the two influences retention.

To determine if study and testing effects vary as a function of test type or item complexity.

To examine the ratings of instruction made by students asked to study for a test as compared with the ratings made by students asked to study "in order to learn rather than for a test."

Method

Subjects

Subjects for this study were 90 undergraduate students enrolled in five educational psychology classes at a large public university which attracts students with diverse backgrounds from an extensive geographical area. Included were both female (N = 60) and male (N = 30) education (N = 62) and noneducation (N = 28) majors (median age = 20 years).

Procedure

Treatment. Treatment in this experiment consisted in part of two different study conditions: test and no test. In the test condition, subjects read the text assignment and attended class with the expectation of being tested. In the no test condition, subjects were asked to "read the text assignment and attend class to learn rather than for a test."

Treatment in this study further consisted of two different kinds of tests, multiple-choice (30 items) and short answer (30 items created from the multiple-choice item stems), each containing two different levels of item complexity,

knowledge (15 items) and concept (15 items). Using the class textbook as a content base for the construction of these tests, the experimenters, who were themselves university professors knowledgeable in the area of educational psychology, first wrote an initial pool of multiple-choice test items at two levels of complexity, (a) knowledge and (b) concept, following the guidelines given by Jenkins and Deno (1971). From this initial pool of items which bore no designation as to level, one of the experimenters selected the 15 items which in her judgment were the best knowledge items and the 15 items which were the best concept items. The resulting 30 items were given to a second researcher who independently labeled each item as being one which measured knowledge or concepts. The raters were in agreement on all items except two. Minor revisions were made in these two items to effect 100% interrater agreement on the knowledge-concept ratings for the items which were than randomly assigned their order of appearance on the test.

Each multiple-choice item had been written so that the stem alone would form a short-answer test item. The short-answer test for this experiment was thus composed of the stems of the multiple-choice test items with the alternatives omitted. Knowledge and concept levels of the short-answer items therefore corresponded to the knowledge and concept levels of the respective multiple-choice items, and the random order of appearance which was used on the multiple-choice test was followed on the short-answer test.

Experimental procedures. Using class rolls, subjects within each class were randomly assigned to one of six groups and each group was randomly assigned (a) study condition and (b) test treatment as follows:

Group 1: (a) test, (b) multiple-choice;

Group 2: (a) test, (b) short answer;

Group 3: (a) test, (b) no test;

Group 4: (a) no test, (b) multiple-choice;

Group 5: (a) no test, (b) short answer;

Group 6: (a) no test, (b) no test.

During the introductory meeting of each of the educational psychology classes students were given by their instructor a syllabus with assignments and test dates (study condition-test: Groups 1, 2, 3). Each instructor subsequently called the names of those students who had been assigned to study condition-no test (Groups 4, 5, 6) and asked them to stay briefly after class where they met with an experimenter who explained that they had been selected to participate in an evaluation of ongoing instructional methods. For their participation, which they were not to discuss with anyone, they would be given an "A" in lieu of their earned grade on the upcoming test. They were, however, still asked to "read the textbook assignment and attend class but in order to learn rather than for a test."

Each of three instructors (white females) taught behavioristic learning theory to her respective class(es) using a lecture-discussion approach for the next 2 weeks. On the assigned test day one of the experimenters, introduced as the departmental coordinator of instruction, asked all students to complete a 25-item Likert-type rating scale for the unit just completed in order to provide evaluative feedback regarding instruction in these classes. (This rating scale contained in random order 4 items related to instructional method, 4 items related to material studied, 4 items related to student effort, 3 items related to usefulness or relevancy of the material, 3 items related to student level of motivation, 4 items related to student achievement, and 3 general evaluative items.) When all students had made their ratings, the

experimenter asked for by name those students who had been assigned to the study condition-no test group (Groups 4, 5, 6) as well as those in Group 3.

Subjects in Group 3--study condition-test, test condition-no test--and Group 6--study condition-no test, test condition-no test--went with one experimenter to a vacant classroom where they were told, as some of the group already knew, that their class was participating in an evaluation of instructional techniques which would necessitate their not responding to the test their classmates were taking.

Subjects in Group 4--study condition-no test, test condition-multiple-choice--and Group 5--study condition-no test, test condition--short answer--went with another experimenter to a vacant classroom where it was explained to them that one of their functions in the evaluation project previously discussed was to respond to the test their classmates were taking. All agreed to continue to participate and were administered, acccording to their respective group, either a multiple-choice or a short-answer test.

Meanwhile, after explaining that the coordinator of the educational psychology classes was working with the students who had left the room, the instructor in the regular classroom routinely administered from one common stack either a multiple-choice or a short-answer test respectively to students in Group 1--study condition-test, test condition-multiple-choice--and Group 2--study condition-test, test condition-short answer. At the beginning of the next class meeting, students in Groups 1 and 2 were informed, as their classmates had been earlier, that their class was participating in an evaluation of instructional procedures which would be explained further at a later date.

Six weeks later, during which time the regularly scheduled classroom activities ensued, an experimenter came back to each class on an unannounced

basis and administered both experimental tests to all students. They were told that their performance on these tests was the concluding part of the evaluation they had earlier been asked to participate in. Each person was strongly encouraged to do his or her best on both tests with an added incentive for conscientious effort being an "A" instead of the unit test scores.

Data Preparation and Analyses

In order to guard against any possible bias, all identifying information was concealed on the achievement tests, the retention tests, and the evaluation forms. The multiple-choice tests were then scored using an objective scoring key so that both a knowledge score (number of knowledge items correct) and a concept score (number of concept items correct) resulted along with a total score (total number or items correct). A detailed scoring key giving the specific acceptable answers was prepared and used to score the short-answer tests with knowledge, concept, and total scores resulting. On the evaluation scale, item ratings were summed within subsets of items so that scores resulted for method, material, effort, usefulness, motivation, achievement, and general.

Study condition differences in total achievement and in the knowledge and concept measures from both the multiple-choice and the short-answer tests were analyzed using *t* tests.

Resulting scores from the retention tests were analyzed using a 2 X 3 X 2 X 2 factorial analysis of variance with repeated measures on the last two factors. The two between factors were study condition (test, no test) and test treatment condition (multiple-choice, short answer, and no test). The within factors were item type on the retention measure (multiple-choice, short answer) and item complexity within the retention test (knowledge, concept).

For all significant interactions and for appropriate main effects Tuckey's HSD test was used for making pairwise comparisons of the means.

The summed item ratings for method, material, effort, usefulness, motivation, achievement, and general given by those students told to study for a test (study condition-test--Groups 1, 2, and 3) were compared using t tests with ratings given by students asked to study to learn rather than for a test (study condition-no test--Groups 4, 5, and 6).

Results

On the multiple-choice achievement test given immediately following instruction, subjects in the study condition-test group scored significantly higher than subjects in the study condition-no test group on the total measure, t (28) = 3.62, $p < .001$ ($\bar{X}$s = 25.33 and 21.33 respectively), on the knowledge measure, t (28) = 3.07, $p < .01$ ($\bar{X}$s = 12.00 and 9.87 respectively), and on the concept measure, t (28) = 3.41, $p < .01$ ($\bar{X}$s = 13.33 and 11.47 respectively). The same pattern of results was obtained on the short-answer achievement test given immediately following instruction. Subjects in the study condition-test group also scored significantly higher than subjects in the study condition-no test group on the total measure, t (28) = 4.51, $p < .001$ ($\bar{X}$s = 17.80 and 9.07 respectively), on the knowledge measure, t (28) = 3.70, $p < .001$ ($\bar{X}$s = 9.20 and 5.27 respectively), and on the concept measure, t (28) = 4.89, $p < .001$ ($\bar{X}$s = 8.60 and 3.80 respectively).

A summary of the analyses of variance of the effects of study and test treatment conditions on the retention measures considering both item type and item complexity is given in Table 1.

Insert Table 1 About Here

Item type/item complexity retention test means and standard deviations in the study and test treatment conditions are shown in Table 2.

Insert Table 2 About Here

In the analyses of the retention tests, the main effect for study condition was significant, F (1, 84) = 11.65, $p < .001$. Subjects in the study condition-test group ($\bar{X}$ = 8.79) scored higher than subjects in the study condition-no test group ($\bar{X}$ = 7.53). (Note: All means reported are an average of the means for levels of factors involved.) The main effect for the test treatment condition was significant, F (2, 84) = 4.95, $p < .01$. Although differences did exist among the multiple-choice treatment group ($\bar{X}$ = 8.83), the short-answer treatment group ($\bar{X}$ = 8.24), and the no test treatment group ($\bar{X}$ = 7.41), these differences were not explored due to the significant interaction between study condition and test treatment condition, F (2, 84) = 6.50, $p < .01$. These interactions can be be understood by examining Figure 1. Results of Tuckey's

Insert Figure 1 About Here

test revealed that study condition-test subjects who took either the multiple-choice test ($\bar{X}$ = 9.58) or the short-answer test ($\bar{X}$ = 9.62) scored higher than subjects who took no test ($\bar{X}$ = 7.17), but the means for the multiple-choice test group and the short-answer test group did not significantly differ. Within the study condition-no test, however, means for subjects in the multiple-choice ($\bar{X}$ = 8.07), short-answer ($\bar{X}$ = 6.87) and no test ($\bar{X}$ = 7.65) treatment groups did not differ significantly.

Although there was a significant main effect for the first within group factor, item type, F (1, 84) = 328.75, $p < .001$, knowing that students scored higher on multiple-choice items than on short-answer items contributes little or no valuable information. However, the significant interaction between the within group factor of item type (multiple-choice vs. short answer) and the between group factor of test treatment condition (multiple-choice, short answer, and no test) is of importance, F (2, 84) = 3.38, $p < .05$. Tuckey's test showed

Insert Figure 2 About Here

that, on the multiple-choice retention measure, subjects who took the multiple-choice test initially as a treatment (multiple-choice treatment group) ($\bar{X}$ = 11.08) scored higher than subjects who took the short-answer test as a treatment (short-answer test treatment group) ($\bar{X}$ = 9.85) and the subjects who received no test at the initial testing time (no test treatment group) ($\bar{X}$ = 9.58). The short-answer test treatment group failed to differ from the no test treatment group. With the responses to the short-answer questions as the dependent measure, a different pattern was found. Subjects in the no test treatment group ($\bar{X}$ = 5.23) scored lower than subjects in the multiple-choice treatment group ($\bar{X}$ = 6.57) and subjects in the short-answer treatment group ($\bar{X}$ = 6.63). Subjects in the latter two groups failed to differ.

Item type failed to interact with the between group factor of study condition.

With the final within group factor, item complexity, there was a significant main effect, F (1, 84) = 11.07, $p < .001$. Item complexity did not interact with either of the between group factors, test treatment condition and study

condition. There was a significant interaction between the two within group factors, item complexity and item type, F (1, 84) = 59.45, $p < .001$.

Analyses of the unit evaluation ratings revealed that subjects who were told to study for a test rated the method of instruction in the unit lower ($\bar{X}$ = 8.31) than did the subjects who were instructed to study in order to learn rather than for a test ($\bar{X}$ - 9.10), t (88) = 2.31, $p < .05$. Subjects in the test study condition rated their effort higher ($\bar{X}$ = 9.33) than did those in the no test study condition ($\bar{X}$ = 8.56), t (88) = 2.22, $p < .05$. Subjects in the test study condition also rated their achievement level higher than subjects in the no test study condition ($\bar{X}$ = 10.13), t (88) = 2.38, $p < .05$.

The interest by subjects in the material covered was not significantly different in the test study condition ($\bar{X}$ = 8.27) and no test study condition ($\bar{X}$ = 8.56), t (88) = 1.00, $p > .05$. The usefulness of the material learned was not rated differently in the test condition ($\bar{X}$ = 11.71) and the no test condition (11.93), t (88) = .47, $p > .05$. Motivation level was not significantly different in the test study condition ($\bar{X}$ = 8.11) and the no test study condition ($\bar{X}$ = 8.51), t (88) = 1.32, $p > .05$. The general rating in the no test study condition ($\bar{X}$ = 11.22) was not significantly different from the test study condition ($\bar{X}$ = 11.82), t (88) = 1.43, $p > .05$.

Discussion

Testing seemed to have facilitated what Rothkopf (1970) called "mathemagenic behavior"--attending behaviors which give birth to learning. Subjects asked to study for a test scored significantly higher on both the knowledge and concept items from the multiple-choice test as well as the short-answer test. They also felt that they had thoroughly mastered the material for the

unit and learned quite a bit. Being satisfied with their accomplishments, they felt that they deserved an "A" as a grade for the unit.

The test resulted in higher student achievement by both eliciting and sustaining study behaviors. Those asked to study for the test reported that they put significantly more effort into the study of this unit than did those asked simply to study in order to learn rather than for a test. Those studying for a test reported that they worked hard for the unit--they put forth their best effort. These results are in line with results from other studies researching how test or test-like questions function with increased time on task associated with testing (cf. Rothkopf, 1970).

Not only did studying for a test affect learning but also it affected retention. Those students who studied for a test retained more of what they had learned when tested weeks later. Also there was a testing effect on retention. Those students who had been tested earlier scored significantly higher than those not tested on a test weeks later. While these are important effects, they are best understood through a look at the associated study condition X test condition interaction that resulted. This interaction helps us to better understand the effects of testing on retention as it indicates that it is not merely testing alone or studying alone but the unique combination of both studying for and taking a test that results in the retention of learning.

A modification of Frase's (1968) analysis of the role of postquestions in reading might be applied to these results. With postquestions or test questions anticipated, there is an attentive response to the text and a careful reading of it as Rothkopf (1970) held and as was supported by self-reports by students in this study who expected a test. When giv n the test, the

respondents read the questions and answered them. Praise for the correct response, in Frase's model, reinforces the attentive behaviors characteristic of the reading or study period thereby strengthening the probability of their occurrence in similar situations in the future. No praise or other feedback was given following the initial test in this study in order to avoid contamination of the effect of testing. However, students in this study had no doubt been so reinforced for studying for previous tests and expected their study behaviors to again result in reinforcement. The tests could have thus acquired control over study behaviors.

The test could facilitate memory by helping to clarify explicitly what of all the material studied is important and should be retained. The learner then commits to associative memory the test questions and the related answers from the materials studied. The questions in turn, become discriminative cues that, when presented in the future, serve to signal the correct response. While this conjecture as to how testing influences learning and retention is speculative, it does have additional support in the research literature (Bull, 1973; Koran & Koran, 1975).

The finding that students scored higher on multiple-choice tests than on short-answer tests is as expected. Over 50 years ago Remmers (1923), Brinkley (1924), and Ruch and Charles (1928) found that, on two tests covering the same material--one a test requiring recognition of the correct answers and the other a test calling for recall and reconstruction of the answers--students did better on the recognition tests. Perhaps contrary to some expectations, however, is the finding that subjects in our study initially tested with the multiple-choice test were able to perform on a retention test as well as or better than subjects initially tested with a short-answer test. Not only

were those subjects initially tested with a multiple-choice test able to score higher on the multiple-choice retention test than those initially tested with a short-answer test but also they were able to score as well as the short-answer treatment group on the short-answer retention measure. These results are contrary to supposedly definitive studies such as those by Meyer (1934, 1935) resulting in a significant difference favoring essay tests on both immediate and delayed (5 weeks) memory using both essay and recognition tests as criteria. Vallance (1947), however, did not find a difference in retention associated with the use of the two types of measures. The subjects in Meyer's studies were told to specifically prepare for the kind of test they received while the subjects in Vallance's study, as well as the subjects in this study, were not. It has therefore been hypothesized (Balch, 1964) that anticipated test mode is an influential variable in studies indicating that recall tests have a greater influence on retention. However, in a study of the effects of type of examination anticipated on test preparation and performance, Hakstian (19[illegible]) found that the kind of examination expected did not affect amount or type of preparation or actual test performance. Differing results from prior studies of the effect of test type on retention might be attributed to a cognitive level as Fisher et al. (Note 1) suggested. However, both item type and item complexity were included as variables in this study, and results of analyses of the interactive effects of these two factors were not significant. It, therefore, seems that results of studies such as those done by Meyer (1934, 1935) cannot be taken as conclusive evidence of the superiority of recall tests for promoting retention as has often been done (Balch, 1964). Multiple-choice tests may be as effective even when item complexity varies.

Although not directly related to the research hypothesis of this study, it is interesting to note that retention of conceptual information is greater than retention of factual knowledge. Such results are consistent with old (Tyler, 1933) as well as new (Fisher et al., Note 1) research. Comparable results are not always found, though (Halpin, Halpin, & Harrington, Note 5). Further research may provide more definitive information regarding this problem. Type of retention measure probably should also be a variable in these future retention studies since a significant retention item type X item complexity interaction resulted in this study.

Even though studying for and taking a test seem to enhance student learning and retention, a not-so-positive evaluation of instruction may be a side effect of testing. Subjects in this study who were asked to study for a test in a unit on learning theory rated the method of instruction significantly lower than did those asked to study in order to learn rather than for a test. In contrast to those tested, subjects not tested liked the way the course was conducted for the unit--in fact, they thought it was great! Learning to them was fun, and they said that they would like for the rest of the units in the course to be taught like the one they had just completed.

Thus, with regard to educational practice that might be recommended based on this study, could we say: Test if you want students to learn and remember--do not test if you want to be popular?

Reference Notes

1. Fisher, K., Williams, S., & Roth, J. Effects of multiple-choice testing on student learning. Paper presented at the meeting of the American Educational Research Association, Boston, April, 1980.

2. Anderson, R. C., Surber, J. R., Biddle, W. B., Zych, P. M., & Lieberman, C. E. Retention of text information as a function of the nature, timing, and number of quizzes. Paper presented at the meeting of the American Educational Research Association, Chicago, April, 1975.

3. Sanjivamurthy, P. T., & Kumar, V. K. Test mode anticipation and performance in an algebra course. Paper presented at the meeting of the American Psychological Association, New York, September, 1979.

4. McKenzie, G. R., & Schadler, A. M. Effects of three practice modes on attention, test anxiety, and achievement in a classroom association learning task. Paper presented at the meeting of the American Educational Research Association, Boston, April, 1980.

5. Halpin, G., Halpin, G., & Harrington, J. Retention in an actual classroom setting as a function of type and complexity of tests. Paper presented at the meeting of the American Educational Research Association, San Francisco, April, 1979.

References

Balch, J. The influence of the evaluating instrument on students' learning. American Educational Administration and Supervision, 1964, 1, 169-182.

Brinkley, S. G. Values of new type examinations in high school. Contributions to Education, No. 161. New York: Teachers College, Columbia University, 1924.

Bull, S. G. The role of questions in maintaining attention to textual material. Review of Educational Research, 1973, 43, 83-87.

Frase, L. T. Effect of question location, pacing and mode upon retention of prose material. Journal of Educational Psychology, 1968, 59, 244-249.

Frase, L. T. Boundary conditions for mathemagenic behaviors. Review of Educational Research, 1970, 40, 337-348.

Gaynor, J., & Millham, J. Student performance and evaluation under variant teaching and testing methods in a large college course. Journal of Educational Psychology, 1976, 68, 312-317.

Hakstian, A. R. The effects of type of examination anticipated on test preparation and performance. Journal of Educational Research, 1971, 64, 319-324.

Jenkins, J. R., & Deno, S. L. Assessing knowledge of concepts and principles. Journal of Educational Measurement, 1971, 8, 95-101.

Koran, M. L., & Koran, J. J., Jr. Interaction of learner aptitudes with question pacing in learning from prose. Journal of Educational Psychology, 1975, 67, 76-82.

Landauer, T. K., & Ainslie, K. I. Exams and use as preservatives of course-acquired knowledge. Journal of Educational Research, 1975, 69, 99-105.

LaPorte, R. E., & Voss, J. F. Retention of prose materials as a function of post-acquisition testing. Journal of Educational Psychology, 1975, 67, 259-266.

Meyer, G. An experimental study of the old and new types of examination: I. The effect of the examination set on memory. Journal of Educational Psychology, 1934, 25, 641-661.

Meyer, G. An experimental study of the old and new types of examination: II. Methods of study. Journal of Educational Psychology, 1935, 26, 30-40.

Remmers, H. H. An experimental study of the relative difficulty of true-false, multiple-choice, and incomplete-sentence type of examination questions. Journal of Educational Psychology, 1923, 14, 367-372.

Rothkopf, E. Z. The concept of mathemagenic activities. Review of Educational Research, 1970, 40, 325-336.

Rothkopf, E. Z. Variable adjunct question schedules, interpersonal interaction, and incidental learning from written material. Journal of Educational Psychology, 1972, 63, 87-92.

Ruch, G. M., & Charles, J. W. A comparison of five types of objective tests in elementary psychology. Journal of Applied Psychology, 1928, 12, 398-403.

Sanders, J. R. Retention effects of adjunct questions in written and aural discourse. Journal of Educational Psychology, 1973, 65, 181-186.

Tyler, R. W. Permanence of learning. Journal of Higher Education, 1933, 4, 203-204.

Vallance, T. R. A comparison of essay and objective examinations as learning experiences. Journal of Educational Research, 1947, 41, 279-288.

Voss, J. F. Acquisition and nonspecific transfer effects in prose learning as a function of question form. Journal of Educational Psychology, 1974, 66, 736-740.

White, H. B. Testing as an aid to learning. Educational Administration and Supervision, 1932, 18, 41-46.

Table 1

ANOVA of Study Condition, Test Treatment Condition, Item Type, and Item Complexity for the Retention Tests

Source	Sum of Squares	df	F
Study Condition (Test - No Test)	143.14	1	11.65***
Test Condition (Multiplie-Choice, Short Answer, No Test)	121.67	2	4.95**
Study Condition X Test Condition	159.75	2	6.50**
Between Group Error	1031.66	84	
Item Type (Multiple-Choice, Short Answer)	1460.05	1	328.75***
Item Type X Study Condition	7.80	1	1.76
Item Type X Test Condition	30.02	2	3.38*
Item Type X Study Condition X Test Condition	12.29	2	1.38
Within Group Error (1)	373.06	84	
Item Complexity (Knowledge, Concepts)	40.67	1	11.07***
Item Complexity X Study Condition	.62	1	.17
Item Complexity X Test Condition	2.02	2	.28
Item Complexity X Study Condition X Test Condition	13.40	2	1.82
Within Groups Error (2)	308.53	84	
Item Type X Item Complexity	95.07	1	59.45***
Item Type X Item Complexity X Study Condition	1.00	1	.63
Item Type X Item Complexity X Test Condition	5.76	2	1.80
Item Type X Item Complexity X Study Condition X Test Condition	1.09	2	.34
Within Group Error (3)	134.33	84	

*$p < .05$.
**$p < .01$.
***$p < .001$.

Table 2

Retention Test Means and Standard Deviation by

Study Condition, Test Treatment, Item Type, and Item Complexity

Study Condition/ Test Treatment	Item Type/Item Complexity							
	Multiple-Choice Knowledge		Multiple-Choice Concept		Short Answer Knowledge		Short Answer Concept	
	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
Test/Multiple-Choice	10.87	1.81	12.47	1.68	7.80	2.60	7.20	3.12
Test/Short Answer	9.60	2.23	12.13	1.96	8.47	1.77	8.27	2.79
Test/No Test	8.67	2.89	10.20	2.37	5.07	1.83	4.73	2.66
No Test/Multiple-Choice	9.60	1.72	11.40	1.76	5.87	2.26	5.40	2.69
No Test/Short Answer	8.27	2.63	9.40	2.16	5.53	1.92	4.27	1.16
No Test/No Test	8.93	2.09	10.53	2.45	5.20	2.54	5.93	3.65

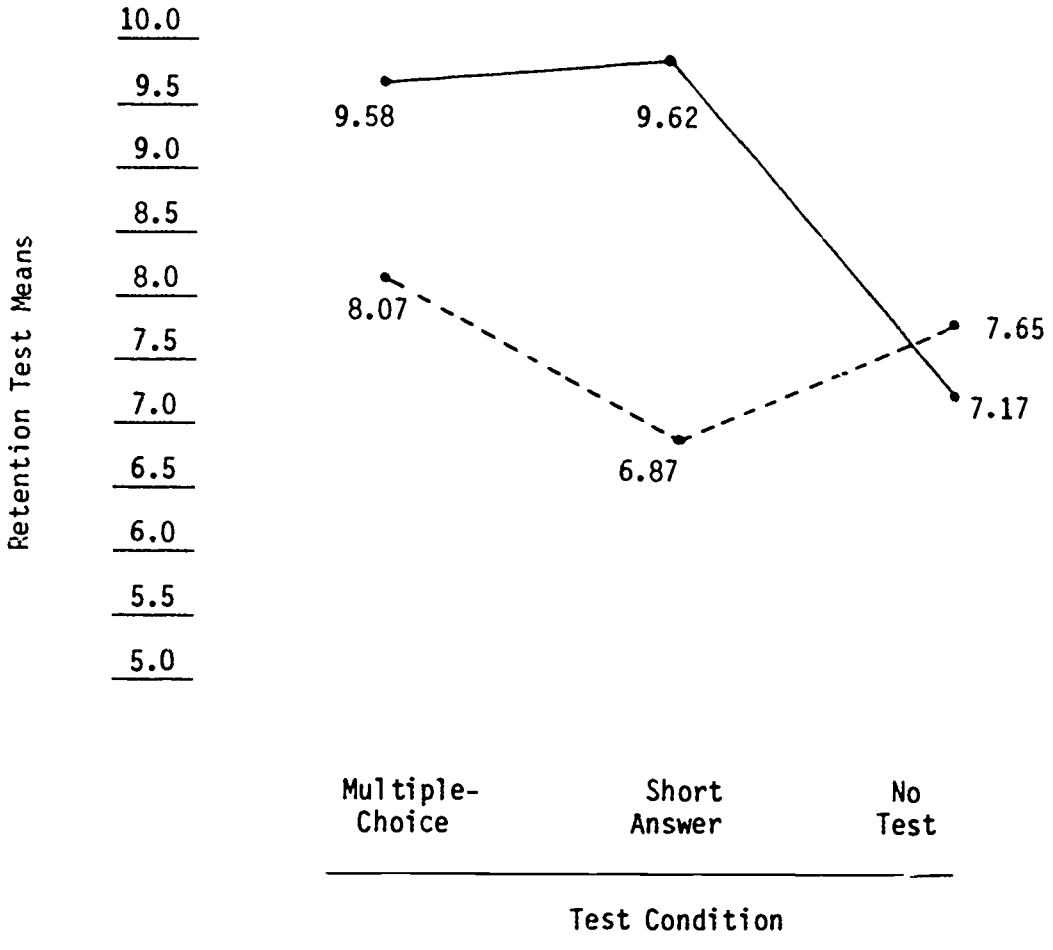

Figure 1. Interaction of test treatment condition and study condition.

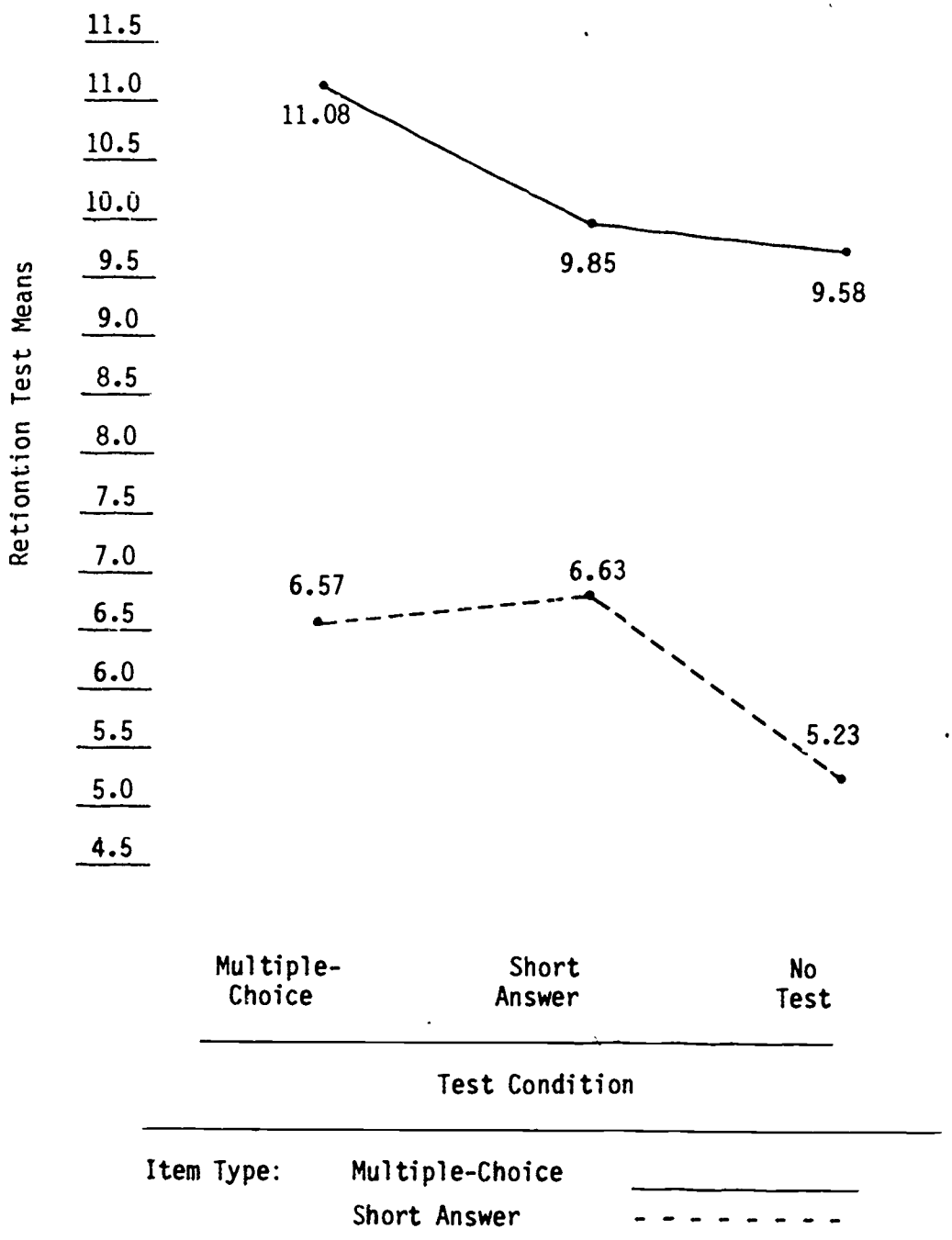

Figure 2. Interaction of item type and test treatment condition.